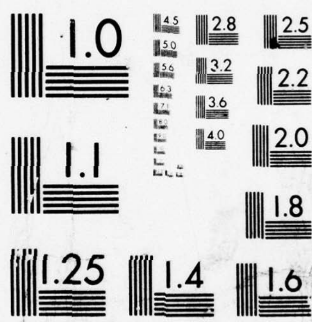
AD-A064 018 OKLAHOMA STATE UNIV STILLWATER DEPT OF STATISTICS F/G 12/1
A STATISTICAL TOOL: ANALYSIS OF COVARIANCE. VOLUME II. THEORETI--ETC(U)
APR 77 S CAUDILL, D HOLBERT, L D BROEMELING F08635-76-C-0154

UNCLASSIFIED AFATL-TR-77-54-VOL-2 NL

1 OF 1
AD
A064018

END
DATE
FILMED
3--79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(18) (19) (2) LEVEL III

AFATL-TR-77-54-VOL-2

A064017

(6)

A STATISTICAL TOOL: ANALYSIS OF COVARIANCE.

VOLUME II. THEORETICAL DEVELOPMENT FOR HANDLING MULTIVARIATE COVARIANCE DATA WITH MISSING VALUES.

AD A064018

(16) 2549

DEPARTMENT OF STATISTICS
OKLAHOMA STATE UNIVERSITY
STILLWATER, OKLAHOMA 74074

(17) Φ4

(11) (12) 42p. (15) FΦ8635-76-C-Φ154

APR[redacted]977

(9)

FINAL REP[redacted]
JAN[redacted]6-DEC[redacted]R 1976

Approved for public release; distribution unlimited

(10) Sam Caudill,
Don Holbert
Lyle D. Broemeling

AIR FORCE ARMAMENT LABORATORY

AIR FORCE SYSTEMS COMMAND • UNITED STATES AIR FORCE

EGLIN AIR FORCE BASE, FLORIDA

79 01 25 045

411 038

JRB

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFATL-TR-77-54, Volume II

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A STATISTICAL TOOL: ANALYSIS OF COVARIANCE VOLUME II. THEORETICAL DEVELOPMENT FOR HANDLING MULTIVARIATE COVARIANCE DATA WITH MISSING VALUES

5. TYPE OF REPORT & PERIOD COVERED
Final Report
January - December 1976

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Sam Caudill
Dr. Don Holbert
Dr. Lyle D. Broemeling

8. CONTRACT OR GRANT NUMBER(s)
F08635-76-C-0154

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Oklahoma State University
Stillwater, Oklahoma 74074

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Program Element 62602F
JON: 2549-04-07

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Armament Laboratory
Armament Development and Test Center
Eglin Air Force Base, Florida 32542

12. REPORT DATE
April 1977

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
Available in DDC

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Analysis of Covariance
Covariance Analysis
Missing Data Routings

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Volume I serves as an introduction to Volumes II and III for those who are not familiar with analysis of covariance. Volume I gives the purpose and uses of analysis of covariance, develops the theory for the univariate cases, expands the theory to the multivariate case, shows how unequal sample size affects the methodology, and how analysis of covariance is used as a tool for evaluating data containing missing observations on the response variable. Section V shows

next page →

ITEM 20 CONCLUDED.

how analysis of covariance may be applied to nonparametric data. Examples depicting each situation are given.

Volume II incorporates all the situations presented in Volume I, except for Section V, and adds the condition of missing observations of the covariate and/or response variables. Volume II presents the theoretical development of the analysis of multivariate covariance in which missing values occur among both dependent and independent variables and presents an example.

Volume III contains the flow chart and program listing of the algorithm developed in Volume II. The theory developed in Volume II will serve for any categorized design, such as a randomized block, Latin square, etc., but the program is only suited for the randomized block model with additive block and treatment effects, or a general two-factor additive effects model with no interaction, or the one-way classification model.

A

PREFACE

This report consists of three volumes which present the theory and application of a valuable data reduction tool, the analysis of covariance. Volume I introduces the analysis of covariance as a general linear model (GLM) and then expands the model to incorporate the multivariate case, unequal sample size, and missing observations on the response variable. Volume I also covers the analysis of covariance for nonparametric data.

Volumes II and III were prepared by the Department of Statistics, Oklahoma State University, Stillwater, Oklahoma 74074, under Air Force Contract F08635-76-C-0154, with the Air Force Armament Laboratory, Armament Development and Test Center, Eglin Air Force Base, Florida 32542. The contract dealt with the development and programming of the methodology for evaluating multiple variable data with missing observations on dependent and independent variables by the analysis of covariance method. The methodology also covers case for unequal sample size. This work was begun in January 1976 and completed in December 1976. This is Volume II.

This technical report has been reviewed by the Information Officer (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

J R Murray

J. R. MURRAY
Chief, Analysis Division

(The reverse of this page is blank)

TABLE OF CONTENTS

(The reverse of this page is blank)

SECTION I

INTRODUCTION AND PROBLEM STATEMENT

The main purpose of this study is to extend work done on estimation and hypothesis testing problems for multivariate linear models describing situations that cannot be analyzed under the Standard Multivariate (SM) general linear model. Kleinbaum (9) has developed the theory to deal with the Growth Curve Multivariate (GCM) model and the More General Linear Multivariate (MGLM) model which is applicable to the problem of missing observations among the dependent variables in the SM model with known design matrix. The author proposes to extend the results of Kleinbaum to handle an analysis of covariance model with missing observations among the independent variables or covariates as well as among the dependent variables.

The Multivariate Analysis of Covariance (MAC) model is based on the multivariate linear model

$$E(Y) = X\alpha + Z\beta \quad \text{and}$$
$$Var(Y) = I_n \otimes \Sigma \qquad (1)$$

where Y is an $n\text{x}p$ matrix composed of p-variate responses on n individuals,

X is an $n\text{x}m_x$ known design matrix of rank $R(X) = r_x(\leq m_x \leq n)$ corresponding to the classificatory variables of the model,

α is an $m_x\text{x}p$ matrix of unknown parameters,

Z is an $n \times m_z$ matrix composed of concomitant variables, in the sense that the constant elements of Z are not necessarily planned in advance by the experimenter. $R(Z) = r_z (\leq m_z \leq n)$,

β is an $m_z \times p$ matrix of unknown concomitant parameters,

$\Sigma = (\sigma_{rs})$ is a $p \times p$ positive definite matrix of usually unknown parameters which represents the variance-covariance matrix of any row of Y,

and $A \otimes B$ is the Kronecker Product of the matrices A and B.

It is clear from Equation (1) that, in the MAC model, the measurements on different individuals are assumed to be uncorrelated whereas the measurements of the p response variates on the same individual may be correlated.

The MAC model may be more concisely represented by using the following definitions:

$A = [X \vdots Z]$ is the $n \times m$ design matrix constructed by horizontally augmenting the design matrix X by the matrix Z where $m = m_x + m_z$,

$\gamma = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ is the $m \times p$ matrix of unknown parameters constructed by vertically augmenting the parameter matrix α by the parameter matrix β.

Thus, the MAC model may be written as follows:

$$E(Y) = A\gamma \text{ and}$$

$$\mathrm{Var}(Y) = I_n \otimes \Sigma. \tag{2}$$

VARIATE-WISE REPRESENTATION OF THE MAC MODEL

The MAC model may be alternatively represented in a variate-wise representation by making the following definitions:

$\mathbf{y}_s$ is the $n \times 1$ vector which denotes the s^{th} $(s = 1,\ldots,p)$ column of Y, and $\underset{\sim}{\gamma}_s$ is mxl vector which denotes the s^{th} $(s = 1,\ldots,p)$ column of γ.

Thus, $Y = [\underset{\sim}{y}_1 \;\; \underset{\sim}{y}_2 \;\; \cdots \;\; \underset{\sim}{y}_p]$

and $\gamma = [\underset{\sim}{\gamma}_1 \;\; \underset{\sim}{\gamma}_2 \;\; \cdots \;\; \underset{\sim}{\gamma}_p]$

so that the MAC model may be described as

$$E(\underset{\sim}{y}_s) = A\underset{\sim}{\gamma}_s, \quad s = 1,2,\ldots, p \text{ and}$$
$$Cov(\underset{\sim}{y}_r,\underset{\sim}{y}_s) = \sigma_{rs}I_n \quad \text{for all} \quad r,s = 1,2,\ldots,p. \tag{3}$$

Thus, the variate-wise representation consists of p univariate models corresponding to the p variates. These p separate **univariate** models are related by the $\frac{p(p-1)}{2}$ covariances between the different variate pairs.

VECTOR REPRESENTATION OF THE MAC MODEL

The vector representation of the MAC Model is obtained by making the following definitions:

$$\text{Let } y = \begin{bmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{y}_p \end{bmatrix} \quad \text{and } \underset{\sim}{\gamma} = \begin{bmatrix} \underset{\sim}{\gamma}_1 \\ \underset{\sim}{\gamma}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{\gamma}_p \end{bmatrix}.$$

Thus,

$$E(\underset{\sim}{y}) = D_A\underset{\sim}{\gamma} \text{ and}$$
$$Var(\underset{\sim}{y}) = \Omega, \tag{4}$$
$$\text{where } D_A = I_p \otimes A \quad \text{and} \quad \Omega = \Sigma \otimes I_n .$$

ESTIMATION AND HYPOTHESES TESTING IN THE MAC MODEL

Rao(12) using generalized inverses has shown for the SM model that the best linear unbiased estimate (BLUE) of a linear function of the elements of the parameter matrix, when estimable, is given by the sum of the BLUE's obtained separately from the univariate models resulting from the variate-wise representation. For estimating an estimable linear set of elements of the parameter matrix, Roy(13) suggests using the sum of the BLUE's for the linear sets obtained separately from the univariate models. The results of Rao and Roy can easily be extended to the MAC model.

The general linear hypothesis for the MAC model can be expressed in the same form as is usual for the SM model for which a number of test procedures have been proposed. For example, Wilk's Likelihood Ratio, Hotelling's Trace (T_0^2), and Roy's Largest Root are the tests most commonly used in practice. Explanations of these tests can be found in standard texts on multivariate analysis such as Anderson(3) and Morrison (10).

EXPERIMENTAL SITUATIONS IN WHICH THE MAC MODEL DOES NOT APPLY

The MAC model, as defined in Equations (1), (2), (3), and (4), involves three assumptions which are not always met in practice due to failure or inability to obtain complete observations on all experimental units.

These assumptions are:

1. A response is observed on each variate on all experimental units.
2. The design matrix, X, is the same for each response variate.
3. Each concomitant response is observed on each experimental unit.

In general, the above assumptions are met in the initial design of an experiment unless it is physically impossible or uneconomical to observe a response on each variate. But even when the experiment is initially designed to conform to the above assumptions, missing observations can occur among the independent as well as the dependent variables due to the occurrence of some unfortunate event such as the dropping of a test tube, the failure of an electronic instrument, or the death of a subject before responses are observed on each variate. These events could be considered random in the sense that their occurrence is equally likely for each experimental unit.

Any failure of the experimental data to conform to the above assumptions yields the MAC model inappropriate for analyzing the experiment based on all observed data, because any experimental units on which one or more dependent and/or independent responses are missing requires the total deletion of that experimental unit. Thus, the development of a procedure utilizing all the sample information would be a valuable contribution to the analysis of such experiments.

SECTION II

LITERATURE REVIEW

Allan and Wishart (1) were probably the first to consider the problem of missing data in statistical analysis, whereas Yates (16) was the first to present a general solution using a least-squares method of substituting for missing values in a designed experiment. Wilks (15) discussed both a maximum likelihood approach and a method-of-moments approach to the problem of missing values in regression analysis.

Zyskind, Kempthorne, et al (17) present a very thorough treatment of the analysis of covariance technique, first introduced by Bartlett (4), to a univariate linear model with missing observations occurring on the dependent variable. They approach the problem by partitioning the model

$$E(\underline{y}) = X\underline{\alpha} \quad \text{and} \qquad \qquad (5)$$
$$\mathrm{Var}(\underline{y}) = \sigma^2 I_n$$

so that it may be written

$$E(\underline{y}) = \binom{X_1}{X_2}\underline{\alpha} \qquad \qquad (6)$$

where $\underline{y}$ is an nx1 vector of observations,

$X = \binom{X_1}{X_2}$ is an nxp known design matrix of full rank $p \leq n$,

and $\underline{\alpha}$ is a px1 vector of unknown parameters.

In general the computational formula for the fitting of a full model of the form [Equation (2)] is used where the data corresponding to the vector $X_1\underline{\alpha}$ of m components are missing or are simply not available. Thus, the model to be fitted is $E(\underline{y}_2) = X_2\underline{\alpha}$, but a solution to the normal equations $X'_2 X_2 \underline{\alpha} = X'_2 \underline{y}_2$ is not immediate, whereas a solution to the normal equations corresponding to the full set of data is standard. They capitalize on the available information by considering the following analysis of covariance model form:

$$E\begin{pmatrix}\underline{0}_m \\ \underline{y}_2\end{pmatrix} = \begin{pmatrix}X_1 \\ X_2\end{pmatrix}\underline{\alpha} + \begin{pmatrix}-I_m \\ 0_{n-m,m}\end{pmatrix}\underline{\beta} \tag{7}$$

where I_m is an mxm identity matrix. Since the sum of squares of deviations of the observations from their expected values for the model [Equation (7)] and the model $E(\underline{y}_2) = X_2\underline{\alpha}$ are minimized for identical sets of values for the vector $\underline{\alpha}$, the computations required for fitting the model $E(\underline{y}_2) = X_2\underline{\alpha}$ can be performed on the corresponding analysis of covariance model. Then using the facts: (i) that for the model

$$E(\underline{y}) = X\underline{\alpha} + Z\underline{\beta} \tag{8}$$

the full set of normal equations

$$X'X\underline{\alpha} + X'Z\underline{\beta} = X'\underline{y} \tag{9}$$

$$Z'X\underline{\alpha} + Z'Z\underline{\beta} = Z'\underline{y} \tag{10}$$

can be equivalently expressed as

$$\begin{aligned} &X'X\underline{\alpha} + X'Z\underline{\beta} = X'\underline{y} \\ &[(I - X(X'X)^{-}X')Z]'\,[(I - X(X'X)^{-}X)Z]\underline{\beta} = [(I - X(X'X)^{-}X')Z]'\,\underline{y} \end{aligned} \tag{11}$$

and (ii) that if $\underline{\lambda}'\underline{\alpha}$ is an estimable parametric function for the model $E(\underline{y}_2) = X_2\underline{\alpha}$ and if for the model $E(\underline{y}) = \begin{pmatrix}\underline{y}_1 \\ \underline{y}_2\end{pmatrix} = \begin{pmatrix}X_1 \\ X_2\end{pmatrix}\underline{\alpha}$ the BLUE of $\underline{\lambda}'\underline{\alpha}$ is given by:

$$\underline{a}_1'\underline{y}_1 + \underline{a}_2'\underline{y}_2\ ; \tag{12}$$

the BLUE of $\underline{\lambda}'\underline{\alpha}$ for the model $E(\underline{y}_2) = X_2\underline{\alpha}$ is given by

$$\underline{a}_1'\hat{\underline{\beta}} + \underline{a}_2'\underline{y}_2 \tag{13}$$

where $\hat{\underline{\beta}}$ is obtained by solving the error normal Equations (11) where

$Z = [-I_m \vdots 0]$ and $\underline{y} = \binom{\underline{0}_m}{\underline{y}_2}$. (13)

Thus, $\hat{\underline{\beta}}$ in Equation (13) plays the role of $\underline{y}_1$ in the point estimation of $\underline{\lambda}'\underline{\alpha}$ for the model $E(\underline{y}) = \binom{\underline{y}_1}{\underline{y}_2} = X\underline{\alpha}$. It would appear that one could easily extend the results of Zyskind, Kempthorne et al to handle the problem of missing responses among the dependent variables of a multivariate linear model. However, this is not the case due to the dependence of their solution upon the fact that the residual sum of squares for the model [Equation (7)] and the model $E(\underline{y}_2) = X_2\underline{\alpha}$ are identical for identical sets of values for the vector $\underline{\alpha}$ which is not guaranteed in the multivariate case due to the covariate structure among responses from the same experimental unit.

Haitovsky (7) compares two alternative methods for dealing with the problem of missing observations among the independent variables and/or the dependent variables in a univariate regression model. One method (Method 1) is simply to discard all incomplete observations and then apply the ordinary least-squares technique to the complete observations. The other method (Method 2) consists of computing the covariances between all pairs of variables, each time using only the observations having values of both variables, and to use these covariances in constructing the system of normal equations.

$Cov(x_i, x_j)\hat{\beta} = Cov(x_i, y)$, $(i,j = 1, \cdots, m)$, where $Cov(x_i x_j)$ (14)
is the mxm covariance matrix in which the $(i,j)^{th}$ element $(i,j = 1,\ldots,m)$ is computed from the measurements common to both x_i and x_j $(i \neq j)$ as well as from all the existing measurements on x_i for $i=j$, and similarly for $Cov(x_i, y)$ $(i = 1,\ldots,m)$. The comparison was made using Monte Carlo techniques since Method 2 does not have optimal statistical properties and since the derivation of its distribution theory is intractable. Comparing the two methods with regard to unbiasedness and efficiency indicated

that Method 2 was superior only in the rare case in which 9 to 10 percent of the observations were complete and hence available for use in Method 1. By decomposing the Mean Square Error (MSE) into one term accounting for bias and the other accounting for the variance when bias is ignored, Haitovsky was able to show that the variance term was far more important in the large difference observed in the two methods. He concluded that, although the bias affects the relevance of the inference, the major problem with Method 2 is caused by the inconsistency introduced into the system of normal Equations (14).

Buck (6) treats the problem of missing values among the dependent variables in a multivariate linear model by estimating the missing values by regression techniques and then calculating a revised variance-covariance matrix. He represents the sample of n experimental units by expressing the responses, y_{ij} ($i = 1,2,\ldots,n$; $j = 1,2,\ldots,p$), in the form of an nxp matrix, Y, in which some of the elements are missing. Assuming that k of the n p-variate responses are complete, he lets these form the first k rows of Y and then calculates the expected value of y_{rj} ($r = 1,2,\ldots,k$) by forming for each value of j, the multiple regression of the j^{th} variable on the other p-1 variables from the set of observations consisting of the first k rows of Y. Thus, he obtains p equations which can be expressed as

$$E(y_{rj}) = f_j(y_{r1}, y_{r2}, \ldots, y_{rj-1}, y_{r,j+1}, \ldots y_{rp}). \qquad (15)$$

The missing values are then estimated as follows. If the i^{th} unit has the j^{th} observation missing, its value, y_{ij}, is estimated by one of the Equations (15) substituting y_{ij} for y_{rj}, that is,

$$E(y_{ij}) = f_j(y_{i1}, y_{i2}, \ldots, y_{ij-1}, y_{ij+1}, \ldots, y_{ip}).$$

This formulation assumes only one missing value in each incomplete response but can be extended to the case in which units have more than one missing value. Buck shows that if the value y_j is missing for a proportion λ of all experimental units, and the predicted values are substituted and a new variance-covariance matrix calculated, then the expectations in this matrix are the same as they would be if there were no missing values, except for the variance ν'_{ij} of y_j which, in terms of expectations, is

$$\nu'_{jj} = \nu_{jj} - \frac{\lambda}{c_{jj}} ,$$

where ν_{jj} is the j^{th} diagonal element of the variance covariance matrix, say V, that would result if there were no missing elements and c_{jj} is the j^{th} diagonal element in V^{-1}.

Beale and Little (5) propose a solution to the problem of missing observations in the dependent variables of a multivariate normal linear model based on the Missing Information Principle of Orchard and Woodbury (11) which involves approximating the Maximum Likelihood solution through an iterative technique. The argument of Beale and Little follows that of Orchard and Woodbury but emphasizes that the effect of the principle is to replace a maximization problem by a fixed point problem. They construct a conditional likelihood function composed of the likelihood equation for known values plus a conditional likelihood of unknown values given the known values and then show that a stationary solution to the conditional likelihood equation is equivalent to the Maximum Likelihood solution based on the original likelihood equation. Thus, assuming the nxp observation matrix, Y, is distributed as a Multivariate Normal, they group the observations into two vectors $\underline{y}$ and $\underline{z}$ with a joint distribution depending on the vector $\underline{\Theta}$ of parameters, where $\underline{y}$ has been observed but $\underline{z}$ has not been observed. To approximate the Maximum Likelihood Estimate (MLE)

$\hat{\underline{\Theta}}$, of $\underline{\Theta}$, based on the log likelihood $L(\underline{y};\ \underline{\Theta})$, they suggest maximizing the expected value of $L(\underline{z},\underline{y};\underline{\Theta})$ where $\underline{z}$ is treated as a random variable with some known distribution. Thus, letting $f(\underline{z}/\underline{y};\underline{\Theta})$ denote the probability density function for the conditional distribution of $\underline{z}$ given $\underline{y}$ and $\underline{\Theta}$, and letting $L(\underline{z}/\underline{y};\underline{\Theta})$ denote $\ln[f(\underline{z}/\underline{y};\underline{\Theta})]$, then

$$L(\underline{z},\underline{y};\underline{\Theta}) = L(y;\Theta) + L(\underline{z}/\underline{y};\underline{\Theta}). \tag{16}$$

A distribution is defined for $\underline{z}$ by taking any assumed value $\underline{\Theta}_A$ for $\underline{\Theta}$ along with the observed value of $\underline{y}$ and one can then take expectations of both sides of Equation (16), integrating with respect to $\underline{z}$. This is expressed by

$$E\{L(\underline{z},\underline{y};\underline{\Theta})/\underline{y};\underline{\Theta}_A\} = L(\underline{y};\underline{\Theta}) + E\{L(\underline{z}/\underline{y};\Theta)|\underline{y};\ \underline{\Theta}_A\}. \tag{17}$$

They then find the value $\underline{\Theta}_M$ of $\underline{\Theta}$ that maximizes the left hand side of Equation (17) and write

$$\underline{\Theta}_M = \emptyset(\underline{\Theta}_A) \tag{18}$$

since $\underline{\Theta}_M$ may depend on $\underline{\Theta}_A$. Thus, Equation (18) represents a transformation from the vector $\underline{\Theta}_A$ to the vector $\underline{\Theta}_M$ from which the Missing Information Principle originates. The Missing Information Principle involves estimating $\underline{\Theta}$ by a fixed point of the transformation, namely a value of $\underline{\Theta}$ such that $\underline{\Theta} = \emptyset(\Theta)$.

As mentioned in the introduction, Kleinbaum (9) proposes a solution to the problem of estimation and hypothesis testing for the MGLM model which is applicable to the case involving missing observations among the dependent variables in the SM model with known design matrix. He writes the SM model in the form

$$E(Y) = X\alpha \quad \text{and} \qquad (19)$$

$$\mathrm{Var}(Y) = I_n \otimes \Sigma$$

where X is an nxm known design matrix of rank $R(X) = r(\leq m \leq n)$,

α is an mxp matrix of unknown parameters, and $\Sigma = ((\sigma_{rs}))$ is a pxp positive definite matrix of usually unknown parameters representing the variance-covariance matrix of any row of Y.

Letting $\underline{y}_s$ be the nx1 vector denoting the s^{th} column of Y and $\underline{\alpha}_s$ the mx1 vector denoting the s^{th} column of α, he writes the variate-wise representation of the SM model as

$$E(\underline{y}_s) = X\underline{\alpha}_s \;, \quad Var(\underline{y}_s) = \sigma_{ss} I_n, \; s = 1, \cdots, p; \qquad (20)$$
$$Cov(\underline{y}_r, \underline{y}_s) = \sigma_{rs} I_n \text{ when } r \neq s.$$

Then stacking the observation vectors on top of one another, the vector representation of the SM model becomes

$$E(\underline{y}) = D_X \underline{\alpha} \qquad (21)$$
$$Var(\underline{y}) = \Omega$$

where $D_X = I_p \otimes X$ and $\Omega = \Sigma \otimes I_n$.

From these representations Kleinbaum develops a general form of the model which allows the omission of responses from variates not observed on a given experimental unit. For the case involving missing observations among the dependent variables of an SM model, he constructs the generalized model as follows. Assuming there are n experimental units and p response variates $V_1, \ldots, V_p$ in total, he lets $\underline{z}_s$, $s = 1, \ldots, p$ be the vector of length N_s, say, corresponding to all observations on V_s in the entire experiment and lets X_s be the $N_s$xm design matrix corresponding to $\underline{z}_s$, i.e., X_s is determined from X by ommitting the rows which correspond to missing values of $\underline{y}_s$. He then lets the N_rxN_s $(r < s)$ matrix Q_{rs} denote the incidence matrix of 0's and 1's defined by $Q_{rs} = (q_{ij(rs)})$ where

$$q_{ij(rs)} = \begin{cases} 1 & \text{if the } i^{th} \text{ component of } \underline{z}_r \text{ and the } j^{th} \text{ component of } \underline{z}_s \text{ are observed on the same experimental unit,} \\ 0 & \text{otherwise} \end{cases}$$

Thus, the variate-wise representation of the MGLM model is given by

$$E(\underline{z}_s) = X_s\underline{\alpha}_s \qquad Var(\underline{z}_s) = \sigma_{ss}I_{N_s} \tag{22}$$

$$Cov(\underline{z}_r,\underline{z}_s) = \sigma_{rs}Q_{rs},\ r < s$$

$$Cov(z_r,z_s) = \sigma_{rs}Q'_{rs},\ r > s,\ r,s = 1,\ldots,p.$$

and with the above definitions the vector representation of the MGLM is given by:

$$E(\underline{z}) = \begin{bmatrix} X_1 & & & & \phi \\ & X_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ \phi & & & & X_p \end{bmatrix} \underline{\alpha} \quad \text{and} \quad Var(\underline{z}) = \Omega \tag{23}$$

where

$$z_{(N\times 1)} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{z}_p \end{bmatrix}, \qquad \underline{\alpha}_{(M\times 1)} = \begin{bmatrix} \underline{\alpha}_1 \\ \underline{\alpha}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\alpha}_p \end{bmatrix},$$

$$\Omega_{(N\times N)} = \begin{bmatrix} \sigma_{11}I_{N_1} & \sigma_{12}Q_{12} & \cdots & \sigma_{1p}Q_{1p} \\ \sigma_{12}Q'_{12} & \sigma_{22}I_{N_2} & \cdots & \sigma_{2p}Q_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1p}Q'_{1p} & \sigma_{2p}Q'_{2p} & \cdots & \sigma_{pp}I_{N_p} \end{bmatrix},$$

$$N = \sum_{s=1}^{p} N_s \quad \text{and} \quad M = \sum_{s=1}^{p} m_s .$$

Kleinbaum then shows that the unique BLUE of any estimable linear function or linear set of the treatment parameters is given by a linear function or linear set, respectively, which involves the unknown parameters of the variance matrix Ω. In fact, restricting linear estimates to be known functions not involving Ω requires additional restrictive conditions on the model. Therefore, he considers Best Asymptotically Normal (BAN) estimation which is a nonlinear method of estimation using estimates of Ω and yielding variances that are, in large samples, the minimum that could be achieved by linear estimators if Ω were known.

For testing linear hypotheses in the MGLM model, assuming the data is normally distributed, Kleinbaum suggests using test statistics which are quadratic forms called Wald Statistics and are constructed from BAN estimators of linear functions of the treatment parameters. Since the asymptotic distribution of a Wald Statistic is a central chi-square variable, the test criteria yield chi-square tests when the sample size is large.

Attempts have been made by several authors to obtain Maximum Likelihood Estimates (MLE) of the parameters in a multivariate linear model with missing observations among the dependent variables. However, most of these methods are applicable to only very specific models. For instance, Anderson (2) describes an iterative technique for obtaining the MLE's of $\alpha = \underline{\alpha}'_{(p\times 1)}$ and Ω when X_s is an $(N_s \times 1)$ vector of ones. Hocking and Smith (8) have developed a procedure for obtaining BAN estimators of α and Ω for the multivariate linear model with missing observations among the dependent variables and they have shown for a special case that their approach yields the maximum likelihood solutions obtained by Anderson. Their estimation procedure involves obtaining initial estimates of the parameters from the group of observations with no missing values and then modifying these initial estimators by adjoining the information in all the remaining groups in a sequential manner by the addition of linear combinations of zero expectations. However, for purposes of a general computer program, extremely cumbersome notation would be required to express the formulae for calculating the estimators at each stage. In fact, Hocking and Smith have only considered a few special cases which involve simply structured models.

SECTION III

PROPOSED SOLUTION

It appears that, if it were possible to generalize the results cited in the literature which deal with missing observations, at best one would have procedures for handling missing values among the dependent and/or independent variables in a univariate analysis of covariance model or missing values among the dependent variables in a multivariate analysis of covariance model. The general form of the SM model for missing observations among the dependent variables, as discussed by Srivastava (14) and Kleinbaum (9), does, however, appear to be valuable as an initial representation of a MAC model in which missing observations occur among the dependent and/or independent variables. In fact, the results of Kleinbaum for estimation and hypothesis testing in the MGLM can be generalized to the More General Multivariate Analysis of Covariance (MGMAC) model by employing a procédure for dealing with the missing independent variables similar to that employed by Zyskind, Kempthorne, et al to deal with missing dependent variables in a univariate linear model.

THE MAC MODEL WITH MISSING DEPENDENT AND/OR MISSING INDEPENDENT VARIABLES (MGMAC)

For purposes of clarity and simplification, the general form of the MGMAC model will be presented by first rewriting the various forms of the MAC model, then generalizing to the General Multivariate Analysis of Covariance (GMAC) model (i.e., the MAC with missing dependent variables), and finally by extending the GMAC to the MGMAC model (i.e., with missing dependent and/or independent variables). To make the presentation as brief as possible, definitions of variables and parameters previously defined will be omitted unless specifically needed for clarification.

The Multivariate Analysis of Covariance Model (MAC) can be represented by

$$E(Y) = X\alpha + Z\beta \qquad (24)$$

$$Var(Y) = I_n \otimes \Sigma$$

or alternatively by

$$E(Y) = A\gamma \qquad \text{where} \quad A = [X \vdots Z] \; . \qquad (25)$$

Thus, the variate-wise representation of the MAC is given by

$$E(\underline{y}_s) = A\underline{\gamma}_s, \quad s = 1,\ldots,p \;\; \text{and} \qquad (26)$$

$$Cov(\underline{y}_r, \underline{y}_s) = \sigma_{rs} I_n \quad \text{for all} \quad r,s, = 1,\ldots,p.$$

and the vector representation is given by

$$E(\underline{y}) = D_A \underline{\gamma} \quad \text{and} \quad Var(\underline{y}) = \Omega \qquad (27)$$

where

$$D_A = I_p \otimes A \quad \text{and} \quad \Omega = \Sigma \otimes I_n.$$

To obtain the general form of the GMAC, assume there are n experimental units and p response variates $V_1,\ldots,V_p$ in total. Let $\underline{z}_s$, $s = 1,\ldots,p$ be the vector of length N_s, say, corresponding to all observations on V_s in the entire experiment. Let $A_{s(N_s x m)}$, $s = 1,\ldots,p$ be the design matrix corresponding to $\underline{z}_s$, i.e., A_s is determined from A by deleting those rows which correspond to missing values of $\underline{y}_s$. Let $Q_{rs(N_r x N_s)}$, $r < s$ denote the incidence matrix of 0's and 1's defined by $Q_{rs} = (q_{ij_{(rs)}})$ where

$$q_{ij_{(rs)}} = \begin{cases} 1 & \text{if the } i^{th} \text{ component of } \underline{y}_r \text{ and the } j^{th} \text{ component of } \\ & \underline{y}_s \text{ are observed on the same experimental unit} \\ 0, & \text{otherwise.} \end{cases}$$

Then the variate-wise representation of the GMAC is given by

$$E(\underline{z}_s) = A_s \underline{\gamma}_s \qquad Var(\underline{z}_s) = \sigma_{ss} I_{N_s} \tag{28}$$

$$Cov(\underline{z}_r, \underline{z}_s) = \sigma_{rs} Q_{rs}, \quad r < s$$

$$Cov(\underline{z}_r, \underline{z}_s) = \sigma_{rs} Q'_{rs}, \quad r > s \qquad r,s = 1,\ldots,p.$$

The vector representation of the GMAC is given by

$$E(\underline{z}) = D\underline{\gamma} \quad \text{and} \quad Var(\underline{z}) = \Omega \tag{29}$$

where

$$\underline{z}_{(N\times 1)} = \begin{bmatrix} \underline{z}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{z}_p \end{bmatrix}, \quad D_{(N\times M)} = \begin{bmatrix} A_1 & & & \phi \\ & A_2 & & \\ & & \ddots & \\ \phi & & & A_p \end{bmatrix}, \quad \underline{\gamma}_{(M\times 1)} = \begin{bmatrix} \underline{\gamma}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\gamma}_p \end{bmatrix}$$

$$\Omega_{(N\times N)} = \begin{bmatrix} \sigma_{11} I_{N_1} & \sigma_{12} Q_{12} & \cdots & \sigma_{1p} Q_{1p} \\ \sigma_{12} Q'_{12} & \sigma_{22} I_{N_2} & \cdots & \sigma_{2p} Q_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{1p} Q'_{1p} & \sigma_{2p} Q'_{2p} & \cdots & \sigma_{pp} I_{N_p} \end{bmatrix}$$

$$N = \sum_{s=1}^{p} N_s \quad \text{and} \quad M = mp.$$

To obtain the general form of the MGMAC, assume that the design matrix $A = [X \vdots Z]$ of the MAC model has ${}_{\ell}t$ missing observations in the ℓ^{th} column, $(\ell = m_x+1, \cdots, m_x+m_z)$. Then in the design matrices $A_{s(N_x \times m)}$ of the variate-wise representation of the GMAC model the ℓ^{th} column will have ${}_{\ell}t_s = {}_{\ell}t - k_s$, where k_s is the number of experimental units in which both the independent variable in column ℓ of A_s and the dependent variable on variate V_s are missing. Thus, A_s would have

$t_s = \sum_{\ell=1}^{m} {}_\ell t_s$ missing values.

Then replace A_s by F_s where F_s is derived from A_s by augmenting A_s (with o's in place of missing values) by a matrix A_s^* of dimension $(N_s x t_s)$ composed of t_s columns each with a one in the row position corresponding to the missing values in A_s and zeros elsewhere. [Note: F_s has dimension $(N_s x m_s)$ where $m_s = m + t_s$.] Thus, the variate-wise representation of the MGMAC is given by

$$E(\underline{z}_s) = F_s \underline{\xi}_s, \qquad Var(\underline{z}_s) = \sigma_{rs} I_{N_s} \tag{30}$$

$$Cov(\underline{z}_r, \underline{z}_s) = \sigma_{rs} Q_{rs}, \; r < s$$

$$Cov(\underline{z}_r, \underline{z}_s) = \sigma_{rs} Q'_{rs}, \quad r < s, \; s = 1, \ldots, p$$

where $\underline{\xi}_s = \begin{bmatrix} \underline{\gamma}_s \\ \underline{\delta}_s \end{bmatrix}$ and where $\underline{\delta}_s$ is a $(t_s x 1)$ vector of unknown parameters due to the missing values in A_s.

The vector representation of the MGMAC model is given by:

$$E(\underline{z}) = F\underline{\xi} \quad \text{and} \quad Var(\underline{z}) = \Omega \tag{31}$$

$$\text{where } F_{(N x M)} = \begin{bmatrix} F_1 & & & \phi \\ & F_2 & & \\ & & \ddots & \\ \phi & & & F_p \end{bmatrix}, \quad \underline{\xi} = \begin{bmatrix} \underline{\xi}_1 \\ \underline{\xi}_2 \\ \vdots \\ \underline{\xi}_p \end{bmatrix},$$

$$N = \sum_{s=1}^{p} N_s \quad \text{and} \quad M = \sum_{s=1}^{p} m_s .$$

ESTIMATION FOR THE MGMAC MODEL

Theorem 1: If $\underline{\Theta} = H'\underline{\xi} = \sum_{s=1}^{p} C_s' \underline{\xi}_s$ is estimable, and if Σ is known then $H'\underline{\xi}$ has a unique BLUE given by

$$\hat{\underline{\Theta}} = H'\hat{\underline{\xi}} = H'(F'\Omega^{-1}F)^{-}\ F'\Omega^{-1}\underline{z}$$

whose variance-covariance matrix is given by

$$\mathrm{Var}(\hat{\underline{\Theta}}) = H'(F'\Omega^{-1}F)^{-}H$$

where C_s is a known $(m_s x1)$ vector $(s = 1,\ldots, p)$.

If we **restrict** our estimators to be known linear functions of $\underline{z}$, then we cannot use the $\hat{\underline{\Theta}}$ above unless it is independent of Ω .

Theorem 2: **For the MGMAC model $\hat{\underline{\Theta}}$ is not independent of Ω unless the following conditions are satisfied:**

$$C'_s(F'_sF_s)^{-}F'_sQ_{rs} \in V(F'_s), \quad r < s \quad \text{and}$$

$$C'_s(F'_sF_s)^{-}F'_sQ_{sr} \in V(F'_s), \quad r > s \quad \text{where}$$

$$Q_{rs}(N_r x N_s), \quad r < s \qquad (r,s = 1,\ldots, p)$$

is defined as before.

If the above conditions are not satisfied, one is lead to consider nonlinear methods of estimation which use estimates of Σ and which give variances that are, in large samples, the minimum that could be achieved by linear estimators if Σ were known.

Theorem 3: A BAN estimator which is unbiased for any estimable set $\underline{\Theta} = H'\underline{\xi}$, is given by

$$\hat{\underline{\Theta}}_n = H'\hat{\underline{\xi}} = H'(F'\hat{\Omega}^{-1}F)^{-}F'\hat{\Omega}^{-1}\underline{z} \tag{32}$$

where $\hat{\Omega}$ is obtained from Ω by **substituting the elements of $\hat{\Sigma} = (\hat{\sigma}_{rs})$ given in Theorem 4 below for the corresponding elements of $\Sigma = (\sigma_{rs})$.**

Theorem 4: For the MGMAC model, a consistent and unbiased estimate of Σ is given by $\hat{\Sigma} = (\hat{\sigma}_{rs})$ where

$$\sigma_{ss} = \frac{1}{N_s - R(F_s)}\ \underline{z}'_s\left[I_{N_s} - F_s(F'_sF_s)^{-}F'_s\right]\underline{z}_s, \quad s = 1,\ldots,p \tag{33}$$

and

$$\hat{\sigma}_{rs} = \frac{1}{N_{rs}-R(F_{rs})} \; z'_{rs} \left[I_{N_{rs}} - F_{rs}(F'_{rs}F_{rs})^{-}F'_{rs}\right] \underline{z}_{sr} \;, \; r \neq s$$

$(r,s = 1,2,\ldots,p)$,

where $N_r(\geq 2)$ is the number of experimental units on which V_r is observed,

$N_{rs}(\geq 2)$ is the number of experimental units on which both V_r and V_s are observed together,

$\underline{z}_{rs(N_r x1)}$ is the vector of all observations on V_r,

$\underline{z}_{rs(N_{rs} x1)}$, $r \neq s$ is the vector of observations on V_r which correspond to units on which both V_r and V_s are observed together,

$F_{r(N_r x m_r)}$ is the design matrix corresponding to $\underline{z}_r$, and

$F_{rs(N_{rs} x m_r)}$ is the design matrix corresponding to $\underline{z}_{rs}$.

The proof of the above theorem follows easily from Kleinbaum (9).

Theorem 5: **For the MGMAC model, the asymptotic variance matrix of any BAN** estimator of an estimable linear set $H'\underline{\xi}$, where $H_{(Mxw)}$ is of full rank w, is given by

$$H'(F'\Omega^{-1}F)^{-}H \; .$$

Note: $H'(F'\Omega^{-1}F)^{-}H$ is the same as the variance matrix of the unique BLUE set $\hat{\underline{\theta}} = H'\hat{\underline{\xi}}$ for $H'\underline{\xi}$ when Ω is known.

TESTING LINEAR HYPOTHESES FOR THE MGMAC MODEL

Theorem 6: **For the MGMAC model, let $H'\underline{\xi}$ be estimable where $H_{(Mxw)}$ is known and of full rank w. Then, if the null hypothesis is H_0: $H'\underline{\xi} = 0$,**

$$W_n = (H'\hat{\underline{\xi}})' \left[H'(F'\hat{\Omega}^{-1}F)^{-}H\right]^{-1} (H'\hat{\underline{\xi}}) \tag{34}$$

is asymptotically distributed as a central chi-square variable with w degrees of freedom, where

$$\hat{\Omega} = \Omega \Big|_{\Sigma = \hat{\Sigma}} \;,$$

$\hat{\Sigma}$ is any positive definite consistent estimator of Σ, Ω and F are defined by the vector representation and $\widehat{H'\underline{\xi}}$ is any BAN estimator of $H'\underline{\xi}$. This result is easily extended from Kleinbaum (9).

To test the hypothesis H_0: $H'\underline{\xi} = 0$, we may thus reject H_0 if $W_n \geq \chi^2_{w,1-\alpha}$ and accept otherwise.

NOTE: All the above theorems follow easily from similar theorems by Kleinbaum (9).

SECTION IV

AN EXAMPLE

The following example is given to illustrate the procedures outlined in Section III for testing hypotheses and for obtaining parameter estimates. The data consists of a portion of data from an exercise in Morrison (10). The saml1 sample size was chosen only in order to make the problem manageable for hand computations. The dependent variables represent two characteristics of urine specimens of young men classified into two groups according to their degree of obesity. One measure, specific gravity, was selected as a concomitant variable. The observations on these variates and the concomitant variable are given below (blank spaces represent missing observations):

Group I			Group II		
Y_1	Y_2	X	Y_1	Y_2	X
17.6	5.15	24	18.1	9.00	31
13.4	5.75	32	19.7	5.30	
20.3	4.35	17	16.9	9.85	32
22.3	7.55	30	23.7	3.60	20
20.5	8.50	30	19.2		18
18.5			18.0	4.40	23
12.1	5.95	25	14.8	7.15	31
12.0	6.30	30	15.6	7.25	28
10.1	5.45	28	16.2	5.30	21
	3.75	24			

If there were no missing observations, the model could be written in the form of Equation (24) or (25). However, since observations are missing from columns of Y and A (or Z), writing the model in the form of Equation (25) results in blanks in the Y and A matrices as shown below:

$$E(Y) = A\gamma \quad \text{and} \tag{35}$$

$$Var(Y) = I_n \otimes \Sigma \quad \text{where}$$

$$Y = \begin{bmatrix} 17.6 & 5.15 \\ 13.4 & 5.75 \\ 20.3 & 4.35 \\ 22.3 & 7.55 \\ 20.5 & 8.50 \\ 18.5 & \text{Blank} \\ 12.1 & 5.95 \\ 12.0 & 6.30 \\ 10.1 & 5.45 \\ \text{Blank} & 3.75 \\ 18.1 & 9.00 \\ 19.7 & 5.30 \\ 16.9 & 9.85 \\ 23.7 & 3.60 \\ 19.2 & \text{Blank} \\ 18.0 & 4.40 \\ 14.8 & 7.15 \\ 15.6 & 7.25 \\ 16.2 & 5.30 \end{bmatrix}, \quad A = [X \vdots Z] = \begin{bmatrix} 1 & 0 & 24 \\ 1 & 0 & 32 \\ 1 & 0 & 17 \\ 1 & 0 & 30 \\ 1 & 0 & 30 \\ 1 & 0 & \text{Blank} \\ 1 & 0 & 25 \\ 1 & 0 & 30 \\ 1 & 0 & 28 \\ 1 & 0 & 24 \\ 0 & 1 & 31 \\ 0 & 1 & \text{Blank} \\ 0 & 1 & 32 \\ 0 & 1 & 20 \\ 0 & 1 & 18 \\ 0 & 1 & 23 \\ 0 & 1 & 31 \\ 0 & 1 & 28 \\ 0 & 1 & 21 \end{bmatrix}$$

$$\gamma = \begin{bmatrix} \alpha \\ \hline \beta \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{21} \\ \alpha_{12} & \alpha_{22} \\ \beta_{11} & \beta_{21} \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

Transforming to the vector representation [Equation (29)] illuminates the problem of missing observations among the dependent variables, but blanks remain among the independent variables as shown below:

$$E(\underline{z}) = D\gamma \quad \text{and} \tag{36}$$

$$Var(\underline{z}) = \Omega \quad \text{where}$$

$$\underline{z} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \end{bmatrix} \quad \text{with } \underline{z}_1 = \begin{bmatrix} 17.6 \\ 13.4 \\ 20.3 \\ 22.3 \\ 20.5 \\ 18.5 \\ 21.1 \\ 12.0 \\ 10.1 \\ 18.1 \\ 19.7 \\ 16.9 \\ 23.7 \\ 19.2 \\ 18.0 \\ 14.8 \\ 15.6 \\ 16.2 \end{bmatrix} \quad \text{and } \underline{z}_2 = \begin{bmatrix} 5.15 \\ 5.75 \\ 4.35 \\ 7.55 \\ 8.50 \\ 5.95 \\ 6.30 \\ 5.45 \\ 3.75 \\ 9.00 \\ 5.30 \\ 9.85 \\ 3.60 \\ 4.40 \\ 7.15 \\ 7.25 \\ 5.30 \end{bmatrix}$$

$$D = \begin{bmatrix} A_1 & \phi \\ \phi & A_2 \end{bmatrix} \quad \text{with } A_1 = \begin{bmatrix} 1 & 0 & 24 \\ 1 & 0 & 32 \\ 1 & 0 & 17 \\ 1 & 0 & 30 \\ 1 & 0 & 30 \\ 1 & 0 & \text{Blank} \\ 1 & 0 & 25 \\ 1 & 0 & 30 \\ 1 & 0 & 28 \\ 0 & 1 & 31 \\ 0 & 1 & \text{Blank} \\ 0 & 1 & 32 \\ 0 & 1 & 20 \\ 0 & 1 & 18 \\ 0 & 1 & 23 \\ 0 & 1 & 31 \\ 0 & 1 & 28 \\ 0 & 1 & 21 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 1 & 0 & 24 \\ 1 & 0 & 32 \\ 1 & 0 & 17 \\ 1 & 0 & 30 \\ 1 & 0 & 30 \\ 1 & 0 & 25 \\ 1 & 0 & 30 \\ 1 & 0 & 28 \\ 1 & 0 & 24 \\ 0 & 1 & 31 \\ 0 & 1 & \text{Blank} \\ 0 & 1 & 32 \\ 0 & 1 & 20 \\ 0 & 1 & 23 \\ 0 & 1 & 31 \\ 0 & 1 & 28 \\ 0 & 1 & 21 \end{bmatrix}$$

$$\underline{\gamma} = \begin{bmatrix} \underline{\gamma}_1 \\ \underline{\gamma}_2 \end{bmatrix} \quad \text{with } \underline{\gamma}_1 = \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \beta_{11} \end{bmatrix} \quad \text{and} \quad \underline{\gamma}_2 = \begin{bmatrix} \alpha_{21} \\ \alpha_{22} \\ \beta_{21} \end{bmatrix}$$

and

$$\Omega = \left[\begin{array}{c|c} \sigma_{11} I_{18} & \sigma_{12} Q_{12} \\ \hline \sigma_{12} Q'_{12} & \sigma_{22} I_{17} \end{array}\right] \quad \text{with}$$

$$Q_{12} = \begin{bmatrix}
1 & & & & & & & & & & & & & & & \\
& 1 & & & & & & & & & & & & & & \\
& & 1 & & & & & & & & & & & & & \\
& & & 1 & & & & & & & & & & & & \\
& & & & 1 & & & & & & & \phi & & & & \\
& & & & & 1 & & & & & & & & & & \\
& & & & & & 1 & & & & & & & & & \\
& & & & & & & 1 & & & & & & & & \\
& & & & & & & & 1 & & & & & & & \\
& & & & & & & & & 1 & & & & & & \\
& & & & & & & & & & 1 & & & & & \\
& & & & & & & & & & & 1 & & & & \\
& & & & & & & & & & & & 1 & & & \\
& & \phi & & & & & & & & & & & 1 & & \\
& & & & & & & & & & & & & & 1 & \\
& & & & & & & & & & & & & & & 1
\end{bmatrix}$$

To transform to the vector representation [Equation (31)] of the MGMAC model, A_1 is replaced by F_1, $\underline{\gamma}_1$ is replaced by $\underline{\xi}_1$, A_2 is replaced by F_2 and $\underline{\gamma}_2$ is replaced by $\underline{\xi}_2$ as shown below:

$$E(\underline{z}) = F\underline{\xi} \quad \text{and} \tag{37}$$

$$\mathrm{Var}(\underline{z}) = \Omega \quad \text{where}$$

$$\underline{z} = \begin{bmatrix} \underline{z}_1 \\ \underline{z}_2 \end{bmatrix} \quad \text{as defined before,}$$

$$F = \begin{bmatrix} F_1 & \phi \\ \phi & F_2 \end{bmatrix} \text{ with } F_1 = \begin{bmatrix}
1 & 0 & 24 & 0 & 0 \\
1 & 0 & 32 & 0 & 0 \\
1 & 0 & 17 & 0 & 0 \\
1 & 0 & 30 & 0 & 0 \\
1 & 0 & 30 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
0 & 0 & 25 & 0 & 0 \\
1 & 0 & 30 & 0 & 0 \\
1 & 0 & 28 & 0 & 0 \\
0 & 1 & 31 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 \\
0 & 1 & 32 & 0 & 0 \\
0 & 1 & 20 & 0 & 0 \\
0 & 1 & 18 & 0 & 0 \\
0 & 1 & 23 & 0 & 0 \\
0 & 1 & 31 & 0 & 0 \\
0 & 1 & 28 & 0 & 0 \\
0 & 1 & 21 & 0 & 0
\end{bmatrix} \text{ and } F_2 = \begin{bmatrix}
1 & 0 & 24 & 0 \\
1 & 0 & 32 & 0 \\
1 & 0 & 17 & 0 \\
1 & 0 & 30 & 0 \\
1 & 0 & 30 & 0 \\
1 & 0 & 25 & 0 \\
1 & 0 & 30 & 0 \\
1 & 0 & 28 & 0 \\
1 & 0 & 24 & 0 \\
0 & 1 & 31 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 32 & 0 \\
0 & 1 & 20 & 0 \\
0 & 1 & 23 & 0 \\
0 & 1 & 31 & 0 \\
0 & 1 & 28 & 0 \\
0 & 1 & 21 & 0
\end{bmatrix}$$

$$\underline{\xi} = \begin{bmatrix} \underline{\xi}_1 \\ \underline{\xi}_2 \end{bmatrix} \quad \text{with} \quad \underline{\xi}_1 = \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \beta_{11} \\ \delta_{11} \\ \delta_{12} \end{bmatrix} \quad \text{and} \quad \underline{\xi}_2 = \begin{bmatrix} \alpha_{21} \\ \alpha_{22} \\ \beta_{21} \\ \delta_{21} \end{bmatrix}$$

and $\Omega = \left[\begin{array}{c|c} \sigma_{11} I_{18} & \sigma_{12} Q_{12} \\ \hline \sigma_{12} Q'_{12} & \sigma_{22} I_{17} \end{array}\right]$ as defined earlier.

<u>Estimation of Σ</u>

Using Theorem 4, a consistent and unbiased estimator

$$\hat{\Sigma} = (\hat{\sigma}_{rs}) \tag{38}$$

of Σ is obtained by letting

$$\hat{\sigma}_{11} = \frac{1}{N_1 - R(F_1)} \underline{z}_1' \left[I_{N_1} - F_1(F_1'F_1)^- F_1'\right] \underline{z}_1 ,$$

$$\hat{\sigma}_{22} = \frac{1}{N_2 - R(F_2)} \underline{z}_2' \left[I_{N_2} - F_2(F_2'F_2)^- F_2'\right] \underline{z}_2 \quad \text{and}$$

$$\hat{\sigma}_{21} = \hat{\sigma}_{12} = \frac{1}{N_{12} R(F_{12})} \underline{z}_{12}' \left[I_{N_{12}} - F_{12}(F_{12}'F_{12})^- F_{12}'\right] \underline{z}_{21} \quad \text{where}$$

$N_1 = 18, \; N_2 = 17, \; N_{12} = N_{21} = 16$

$$\underline{z}_1 = \begin{bmatrix} 17.6 \\ 13.4 \\ 20.3 \\ 22.3 \\ 20.5 \\ 18.5 \\ 12.1 \\ 12.0 \\ 10.1 \\ 18.1 \\ 19.7 \\ 16.9 \\ 23.7 \\ 19.2 \\ 18.0 \\ 14.8 \\ 15.6 \\ 16.2 \end{bmatrix}, \quad \underline{z}_2 = \begin{bmatrix} 5.15 \\ 5.75 \\ 4.35 \\ 7.55 \\ 8.50 \\ 5.95 \\ 6.30 \\ 5.45 \\ 3.75 \\ 9.00 \\ 5.30 \\ 9.85 \\ 3.60 \\ 4.40 \\ 7.15 \\ 7.25 \\ 5.30 \end{bmatrix}, \quad \underline{z}_{12} = \begin{bmatrix} 17.6 \\ 13.4 \\ 20.3 \\ 22.3 \\ 20.5 \\ 12.1 \\ 12.0 \\ 10.1 \\ 18.1 \\ 19.7 \\ 16.9 \\ 23.7 \\ 18.0 \\ 14.8 \\ 15.6 \\ 16.2 \end{bmatrix}, \quad \underline{z}_{21} = \begin{bmatrix} 5.15 \\ 5.75 \\ 4.35 \\ 7.55 \\ 8.50 \\ 5.95 \\ 6.30 \\ 5.45 \\ 9.00 \\ 5.30 \\ 9.85 \\ 3.60 \\ 4.40 \\ 7.15 \\ 7.25 \\ 5.30 \end{bmatrix},$$

F_1 and F_2 are defined as before and

$$F_{12}=F_{21}=\begin{bmatrix} 1 & 0 & 24 & 0 \\ 1 & 0 & 32 & 0 \\ 1 & 0 & 17 & 0 \\ 1 & 0 & 30 & 0 \\ 1 & 0 & 30 & 0 \\ 1 & 0 & 25 & 0 \\ 1 & 0 & 30 & 0 \\ 1 & 0 & 28 & 0 \\ 0 & 1 & 31 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 32 & 0 \\ 0 & 1 & 20 & 0 \\ 0 & 1 & 23 & 0 \\ 0 & 1 & 31 & 0 \\ 0 & 1 & 28 & 0 \\ 0 & 1 & 21 & 0 \end{bmatrix}$$

Substitution of the above values into Equation (38) results in

$$\hat{\Sigma} = \begin{bmatrix} 13.9694 & 1.7376 \\ 1.7376 & 1.3775 \end{bmatrix}.$$

ESTIMATION OF ORIGINAL PARAMETERS

A BAN estimator which is unbiased for $H'\underline{\xi} = \underline{\gamma} = \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \beta_{11} \\ \alpha_{21} \\ \alpha_{22} \\ \beta_{21} \end{bmatrix}$ is given by

$$H'\underline{\hat{\xi}} = H'(F'\hat{\Omega}^{-1}F)^{-}F'\hat{\Omega}^{-1}\underline{z} \qquad \text{where} \tag{39}$$

$\hat{\Omega}$ is obtained from Ω by substituting the elements of $\hat{\Sigma}$ given above for the corresponding elements of Σ ,

$$H' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} F_1 & \phi \\ \phi & F_2 \end{bmatrix} \quad \text{and} \quad \underline{z} = \begin{bmatrix} \underline{z}_1 \\ \hline \underline{z}_2 \end{bmatrix}.$$

Substitution of the appropriate values into Equation (39) yields

$$H'\hat{\underline{\xi}} = \hat{\underline{\gamma}} = \begin{bmatrix} \hat{\alpha}_{11} \\ \hat{\alpha}_{12} \\ \hat{\beta}_{11} \\ \hat{\alpha}_{21} \\ \hat{\alpha}_{22} \\ \hat{\beta}_{21} \end{bmatrix} = \begin{bmatrix} 21.3524 \\ 23.0662 \\ -0.2071 \\ -4.0121 \\ -3.2103 \\ 0.3686 \end{bmatrix}$$

In order to have obtained the parameters introduced into the model by the missing observations among the independent variables in addition to the original parameters, $H = I_9$ would have been used in Equation (39).

HYPOTHESIS TEST - NO OVERALL GROUP EFFECT

The joint null hypothesis of no overall group effect, which can be written

$$H_0:\ H'\underline{\xi} = \begin{bmatrix} \alpha_{11} - \alpha_{12} \\ \alpha_{21} - \alpha_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{where}$$

$$H = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & -1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \underline{\xi} = \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \beta_{11} \\ \delta_{11} \\ \delta_{12} \\ \alpha_{21} \\ \alpha_{22} \\ \beta_{21} \\ \delta_{21} \end{bmatrix},$$

is tested by computing

$$W_n = (H'\hat{\underline{\xi}})' \, [H'(F'\hat{\Omega}^{-1}F)^{-}H]^{-1} \, (H'\hat{\underline{\xi}}) \tag{40}$$

Substitution of the appropriate values into Equation (40) results in $W_n = 3.037$ which yields an observed significance level between 0.1 and 0.25 based on the fact that W_n is asymptotically distributed as a central chi-square with $R(H) = 2$ degrees of freedom. Based on the results of this test, it could be concluded that at the commonly accepted levels of significance there is not sufficient evidence to reject the joint null hypothesis of no difference in the characteristics of urine specimens of young men in Group I and Group II.

APPENDIX A

LIST OF SYMBOLS

$\underline{a}(q \times 1)$	is a $(q \times 1)$ column vector, and $\underline{a}'$ is the corresponding $(1 \times q)$ row vector.
$A(p \times q) = (a_{ij})$	is a $(p \times q)$ matrix with a_{ij} as the element in the i^{th} row and j^{th} column.
$A = (A_{ij})$	is a partitioned matrix in which A_{ij} is the sub-matrix in the i^{th} row and j^{th} column.
$R(A)$	is the rank of the matrix A.
$V(A)$	is the vector space spanned by the rows of A.
A'	is the transpose of A.
$\mathrm{tr}\, A$	is the trace of A.
A^{-1}	is the unique inverse of a square matrix A of full rank.
A^{-}	is any generalized inverse of the matrix A and is defined by $AA^{-}A = A$.
$A \otimes B$	is the Kronecker Product of the matrix A and B defined by $A \otimes B = (a_{ij}B)$ where $A = (a_{ij})$.
I_q	is the identity matrix of order q.
$\underline{o}_p$	is the $(p \times 1)$ vector of zeros.
$0_{p,q}$	is the $(p \times q)$ matrix of zeros.

For $\underline{x} = (x_1,\ldots,x_n)'$ and $\underline{y} = (y_1,\ldots,y_m)'$,

Cov $(\underline{x},\underline{y})$	is the $(n \times m)$ matrix with $\mathrm{Cov}(x_i,y_j)$ in the i^{th} row and j^{th} column;
$\mathrm{Var}(\underline{x})$	is the $(n \times n)$ matrix $\mathrm{Cov}(\underline{x},\underline{x})$.

For $Y(n \times p) = (y_{ij})$, a matrix of random variables,

$E(Y)$	is the $(n \times p)$ matrix of expectations of the elemtents of Y, i.e., $E(Y) = (Ey_{ij})$;
$\mathrm{Var}(Y)$	is the $(np \times np)$ variance-covariance matrix of the $(np \times 1)$ vector defined by putting the rows of Y underneath each other in a long column vector
$\underline{x} \sim N_p(\underline{\mu},\Sigma)$	means that the random variable $\underline{x}$ has a p-variate multinormal distribution with mean vector $\underline{\mu}$ and variance-covariance matrix Σ.
$\Omega^{\frac{1}{2}}$	is the square root of a symmetric matrix Ω defined by $\Omega^{\frac{1}{2}} = C'\Lambda C$ where C is an orthogonal matrix and Λ is a diagonal matrix such that $\Omega = C'\Lambda^2C$.
${}_{\ell}t$	is read as "the variable t with left subscript ℓ".
${}_{\ell}t_s$	is read as "the variable t with left subscript ℓ and right subscript s".

APPENDIX B

GLOSSARY OF TERMS

Given: A p-variate random sample, $Y_{(nxp)}$ of size n from a population with probability density function $f(Y,\Theta)$, $\Theta \in \boldsymbol{\Theta}$ (parameter space, then:

An <u>estimator</u>, T, of the parameter $g(\Theta)$, $\Theta \in \boldsymbol{\Theta}$ is a function of Y whose range contains the range of $g(\Theta)$.

An <u>unbiased estimator</u>, T, of $g(\Theta)$ is one such that

$$E(T) = g(\Theta), \quad \forall \Theta \in \boldsymbol{\Theta} .$$

Denote the class of unbiased estimators of $g(\Theta)$ by U_g.

A <u>Minimum Variance Unbiased Estimator (MVUE)</u> of $g(\Theta)$ is a $T \in U_g$ such that

$$Var(T) \leq Var(T^*), \ \forall T^* \in U_g \text{ and } \Theta \in \boldsymbol{\Theta}.$$

Let V_g be the class of all linear unbiased estimators of $g(\Theta)$.

Then $T \in V_g$ if and only if

(i) $T \in U_g$ and

(ii) $T \in \underline{a}'Y$ where $\underline{a}$ is some constant vector.

A Best Linear Unbiased Estimator (BLUE), T, of $g(\Theta)$ is a $T \in V_g$ such that

$$Var(T) \leq Var(T^*), \ \forall T^* \in V_g \text{ and } \Theta \in \boldsymbol{\Theta}.$$

A sequence of random variables $(Z_n: \ n = 1, 2, \cdots)$ <u>converges in distribution</u> to the random variable Z with distribution function F whenever

$$\lim_{n \to p} F_n(z) = F(z), \text{ for all continuity points}$$

x of F. This is denoted by $Z_n \xrightarrow{d} F$, where F_n is the distribution function of $Z_n (n = 1, 2, \cdots)$.

An estimator $\underset{\sim}{\tilde{\Theta}}_{n(u\times 1)}$ based on a sample of n observations is said to be a Best Asymptotic Normal (BAN) estimator for the parameter $\underset{\sim}{\Theta}_{(u\times 1)} = (\Theta_1, \cdots, \Theta_u)'$ provided

$$\sqrt{n}\ B_n^{\frac{1}{2}} (\underset{\sim}{\tilde{\Theta}}_n - \underset{\sim}{\Theta}_o) \xrightarrow{d} N_u(\underset{\sim}{o}_u, I_u) \text{ where}$$

$$B_{n(u\times u)} = \text{Fisher's Information Matrix}$$

$$= E_{\underset{\sim}{\Theta}_o} \left[-\frac{1}{n} \frac{\partial^2 \log \phi_n}{\partial \underset{\sim}{\Theta}^2} \Bigg|_{\underset{\sim}{\Theta}=\underset{\sim}{\Theta}_o} \right], \text{ where}$$

$\underset{\sim}{\Theta}_o$ is the true value of $\underset{\sim}{\Theta}$,

ϕ_n is the likelihood function for the sample,

$\hat{\Theta}_n$ has asymptotic dispersion matrix $\frac{B_n}{n}$.

BIBLIOGRAPHY

1. Allan, F.E. and Wishart, J. (1930). A Method of Estimating the Yield of a Missing Plot in Field Experimental Work, J. Agric. Sci., 20:399-406.

2. Anderson, T.W. (1957). Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations Are Missing. J. Amer. Statistics Association, 52:200-203.

3. Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis, Wiley and Sons, New York.

4. Bartlett, M.S. (1937). Some Examples of Statistical Methods of Research in Agriculture and Applied Biology, J.R. Statist. Soc.-Suppl., 4:139-183.

5. Beale, E.M.L. and Little, R.J.A. (1975). Missing Values in Multivariate Analysis, J. Roy. Stat. Soc., Series B., 37:129-145.

6. Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer, J. Roy. Soc., Series B., 22:302-306.

7. Haitovsky, Y. (1968). Missing Data in Regression Analysis, J. Roy. Soc., Series B., 30: 67-82.

8. Hocking R. and Smith, W.B. (1968). Estimation of Parameters in the Multivariate Normal Distribution With Missing Observations, J. Am. Stat. Assoc., Volume 63, No. 321, pp. 159-173.

9. Kleinbaum, D.G. Estimation and Hypothesis Testing for Generalized Multivariate Linear Models. University of North Carolina Press: Chapel Hill, North Carolina, 1971.

10. Morrison, D.F. (1967). Multivariate Statistical Methods, McGraw-Hill, New York.

11. Orchard, T. and Woodbury, M.A. (1972). A Missing Information Principal: Theory and Applications, In Proc. 6th Berkeley Symp. Math. Statist. Prob., Volume I, pp. 697-715.

12. Rao, C.R. (1965). Linear Statistical Inference and Its Applications, Wiley and Sons, New York.

13. Roy, S.N. (1964). A Report on Some Results in Simultaneous or Joint Linear (Point) Estimation. Institute of Statistics, Mimeo Series No. 394, University of North Carolina, Chapel Hill, N.C.

14. Srivastava, J.N. (1968). On a General Class of Designs for Multiresponse Experiments, Ann. Math. Statist., 39: 1825-1843.

15. Wilks, S.S. (1932). Moments and Distributions of Estimates of Population Parameters From Fragmentary Samples, Ann. Math. Stat., 3:163-195.

16. Yates, F. (1933). The Analysis of Replicated Experiments When the Field Results Are Incomplete, The Empire J. Experimental Agric., 1:129-142.

17. Zyskind, G. and Kempthorne, O., et. al. (1964). Research on Analysis of Variance and Related Topics, Aerospace Research Laboratories, ARL 64-193.

INITIAL DISTRIBUTION

HQ USAF/SAMI	1
USARE/DOQ	1
PACAF/DOOFQ	1
TAC/DRA	1
ASD/ENFEA	1
AUL/LSE-71-249	1
SAC/NRI (STINFO LIB)	1
NWC/CODE 318	1
NWC/CODE 317	1
OO-ALC/MMWMP	2
AFIS/INTA	1
DDC	2
AFATL/DLODL	2
AFATL/DL	1
AFATL/DLY	1
ADTC/XRS	1
AFATL/DLYV	20
AFATL/DLYW	10
USA ENG WatWay Ex Sta/VMS	1
BAL RESRCH LAB/AMXBR-VL	1
AMSAA/DRXSY-J	2
USAMSAA/DRXSY-S	1
ARRADCOM/DRDAR-LCU-TM	1
AFOSR/NM	1
ARMY RESEARCH OFFICE/NC	1
OKLAHOMA ST UNIV/Dept of Stat	20
TAC/INAT	1
USA TRADOC SYS ANN ACT	1
ASD/XRP	1
COMIPAC/I-232	1

(The reverse of this page is blank)